

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau



(43) International Publication Date  
7 October 2004 (07.10.2004)

PCT

(10) International Publication Number  
**WO 2004/086255 A1**

(51) International Patent Classification<sup>7</sup>: **G06F 17/30**

(74) Agent: LLOYD, Barry, George, William; BT Group Legal Intellectual Property Department, PP C5A, BT Centre, 81 Newgate Street, London, Greater London EC1A 7AJ (GB).

(21) International Application Number:

PCT/GB2004/000744

(22) International Filing Date: 24 February 2004 (24.02.2004)

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

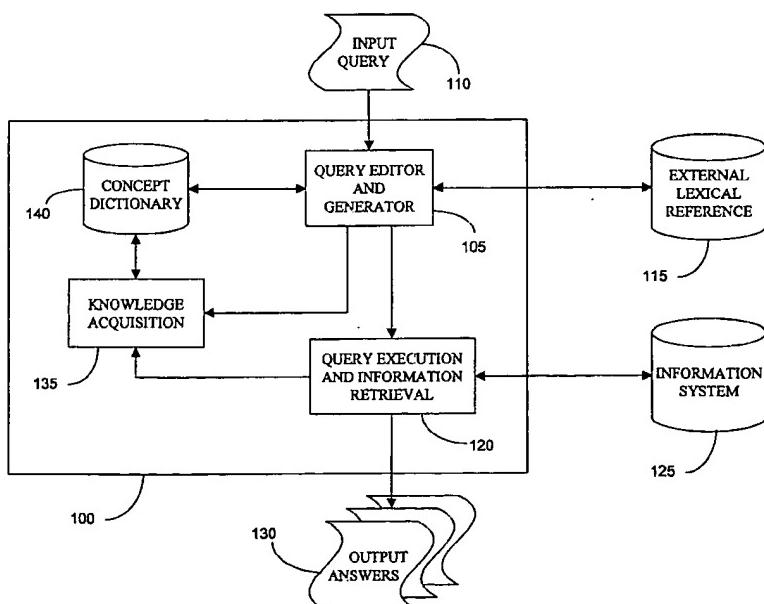
0306877.2 25 March 2003 (25.03.2003) GB

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

*[Continued on next page]*

---

(54) Title: CONCEPT DICTIONARY BASED INFORMATION RETRIEVAL



(57) Abstract: A method and apparatus are provided for generating and updating a concept dictionary (140) in respect of an information system (125) and for using that concept dictionary to assist in selecting queries and query terms for use in interrogating that information system (125). A lexical reference source (115) is first used to generate queries semantically related to a query (110) entered by a user, and the answers returned for each query are analysed using a fuzzy processing technique (135) to determine semantic relationships between the queries. The queries and the determined relationships are recorded in a concept dictionary (140) for subsequent use.

WO 2004/086255 A1



**Published:**

— *with international search report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

3 / PRfs .

10/549365

1 JC05 Rec'd PCT/PTO 14 SEP 2005

## CONCEPT DICTIONARY BASED INFORMATION RETRIEVAL

This invention relates to information retrieval and in particular to a method and apparatus for generating a concept dictionary in respect of an information system for  
5 use in retrieving information from that system.

It is often assumed in prior art electronic information access systems that a user understands something of the structure of the stored data and the methods used to access those data to be able to access relevant information efficiently. In particular, the user may be expected to know terms that appear in stored entries of potential interest and  
10 be able to choose query terms that distinguish these entries from others stored in the system. To help avoid this dependence on user knowledge, it is known to use a thesaurus or ontology to convert queries expressed in the user's favoured terms into queries that may enable the system to retrieve the most relevant entries. For example, if no entries are found in response to a query including the word *pizza*, an ontology might suggest use of  
15 the term *Italian restaurant* instead. However, to be generally applicable, such an ontology must, of necessity, be extremely broad. Experience of the Artificial Intelligence (AI) community suggests that this approach is impractical and that it may be impossible to implement a "universal" AI-based ontology containing human-level general knowledge. In addition, because a universal ontology has to be extremely broad, it tends to over-  
20 generalise queries. For example, the word *car* might be replaced by reasonable synonyms such as *auto*, *automobile*, or *motorcar* but it might also be replaced by *machine*, *railway car*, *elevator car* or *gondola* which are not relevant to the query.

According to a first aspect of the present invention there is provided a method of generating a concept dictionary for use in querying an information system, comprising the  
25 steps of:

- (i) receiving an information search criterion;
- (ii) deriving, using a lexical reference source, at least one search criterion having related meaning to said received search criterion;
- (iii) identifying sets of information in said information system relevant to said received  
30 search criterion and to said at least one derived search criterion;
- (iv) analysing the identified sets of information to derive relationships between said received search criterion and said at least one derived search criterion in the context of said information system; and

(v) storing, in a concept dictionary, information relating to said received and said at least one derived search criterion and to respective said derived relationships therebetween, for use in querying said information system.

The method according to this first aspect of the present invention is particularly applicable to a small subsystem such as an intranet or database, being arranged to deduce the important concepts and their relationships in that limited domain. A local, system-specific concept dictionary or ontology can be used to help a user to generalise, specialise or select equivalent queries and query terms for use in subsequent information retrieval activities without the user becoming lost in over-generalisation.

Recognising that universal ontologies are too general to be of use for query expansion in a relatively limited domain, preferred embodiments of the present invention attempt to extract only that subset of ontological information relevant to the query mechanism and the stored data in a specific information system and to store that ontological information in a concept dictionary specifically relevant to that information system. The concept dictionary is derived with respect to the complete information system, and is not simply a property of the stored data. Interactions between the actual data stored and the mechanism used to access the data have been found to be important to understanding the relationships between queries; relationships that cannot be accurately derived from the stored data alone.

Preferably the concept dictionary is "fuzzy" in that it allows a concept to be approximately equivalent to another concept, or to have partial membership in a parent concept. Fuzzy modelling and processing techniques are described for example in "Fuzzy Sets" by L. Zadeh, Journal of Information and Control, Volume 8, 1965, pp338-353, and "Fuzzy Logic Controllers", Parts 1 and 2, by C. Lee, IEEE Transactions on Systems, Management and Cybernetics, Volume 20, 1990, pp404-435. The application of fuzzy modelling techniques to relate concepts in preferred embodiments of the present invention has been found to be particularly advantageous. Consider, for example, a classified telephone directory. Those directory entries retrieved in response to a query term "garage" might include almost all directory entries that offer "car repair". From this it may be deduced that "car repair" is almost always a more specific concept than "garage". However, relationships derived in this way cannot be guaranteed to be true in all cases. While a conditional probability might be used to relate entries if entries were retrieved with complete certainty, in almost all query-answering systems dealing with semi- or un-structured data different entries satisfy the query to a greater or lesser

degree. Since this degree of satisfaction cannot be treated as a pure probability, it is not possible to apply standard probability theory to the relation between two concepts. However, by treating rankings of entries as fuzzy memberships, uncertain relationships between queries can be modelled, for example relationships such as "car repair is  
5 almost always a more specific query term than garage".

According to a second aspect of the present invention there is provided a method of accessing sets of information stored in an information system using information search criteria stored in a concept dictionary generated for the information system according to the method defined according to the first aspect of the present invention above,  
10 comprising the steps of:

- (a) selecting a first information search criterion;
- (b) using a search engine to identify one or more sets of information in the information system relevant to said first information search criterion; and
- (c) selecting at least one further information search criterion from search criteria  
15 stored in the concept dictionary, semantically related to said first information search criterion according to information stored in the concept dictionary, according to whether a more general, a more specialised or an equivalent search is required.

According to a third aspect of the present invention there is provided an information retrieval apparatus for accessing sets of information stored in an information  
20 system, comprising:

- an input for receiving an information search criterion;
- deriving means for deriving, using a lexical reference source, at least one search criterion having related meaning to said received information search criterion;
- retrieval means for identifying sets of information in said information system  
25 relevant to said received search criterion and to said at least one derived search criterion;
- analysis means for analysing said identified sets of information to derive relationships between said received search criterion and said at least one derived search criterion in the context of said information system; and
- updating means for storing, in a concept dictionary, information relating to said  
30 received and said at least one derived search criterion and to respective said derived relationships therebetween, for use in querying said information system.

Preferred embodiments of the present invention will now be described in more detail, by way of example only, with reference to the accompanying drawings of which:

Figure 1 is diagram showing features of an information retrieval apparatus  
35 according to a preferred embodiment of the present invention;

Figure 2 is flow diagram showing preferred steps in operation of the apparatus of Figure 1; and

Figure 3 is a diagram representing in graphical form an example of knowledge stored in a concept dictionary generated according to preferred embodiments of the 5 present invention.

An apparatus, according to preferred embodiments of the present invention, for use in retrieving information data sets from an information system, will firstly be described with reference to Figure 1.

Referring to Figure 1, a preferred information retrieval apparatus 100 10 comprises a query editor and generator 105 arranged to receive an input query 110 entered by a user or otherwise retrieved from a store of queries. The query editor and generator 105 is arranged with access to an external lexical reference source 115 to enable one or more queries having a related meaning to the input query 110 to be derived, for example by substituting a noun occurring in the input query 110 with a 15 semantically related noun or phrase obtained from the external lexical reference source 115. A lexical database suitable for this purpose is Wordnet™, accessible over the Internet at <http://www.cogsci.princeton.edu/~wn/>.

A query execution and information retrieval module 120 arranged to receive the input query 110 and each of the derived queries generated by the query editor and 20 generator 105, and to identify information data sets, stored in an information system 125, relevant to each of the received queries. The module 120 may be a conventional search engine arranged to implement a known information searching algorithm, preferably one arranged to calculate, for each identified set of information, a weighting factor indicative of the degree of relevance of each identified set of information to the 25 respective executed query.

Those sets of information identified by the information retrieval module 120 as being relevant to the input query 110 in particular, are output as answers 130. In addition the results of the information retrieval (120) in respect of the input query 110 and each of the queries derived by the query editor and generator 105 are received by 30 a knowledge acquisition module 135, together with the input (110) and derived (105) queries themselves, for analysis. The knowledge acquisition module 135 is arranged to execute an algorithm for deriving semantic relationships between the input (110) and derived (105) queries on the basis of the results of information retrieval by the module 120 from the information system 125. In particular, the knowledge acquisition module 35 135 is arranged to determine whether one of the queries, or terms comprised in the

query, represents a specialisation or a generalisation of another of the queries on the basis of the relative scope of information retrieved by the module 120. In this way, any semantic relationships suggested with reference to the external lexical reference 115 when generating the derived queries (105) are tested in the specific context of the  
5 information system 125 and a measure of the extent to which the suggested relationships apply in that context is determined by the knowledge acquisition module 135. A store is provided to store a concept dictionary 140 in respect of the information system 125, the concept dictionary 140 comprising a record of each of the queries, input (110) and derived (105), and the respective measures determined by the  
10 knowledge acquisition module 135 of semantic relationships therebetween, or between terms comprised in those queries. As new input queries 110 are received, the knowledge acquisition module 135 is able to update the concept dictionary 140 by adding new queries and new relationships and by updating values associated with previously stored relationships, thereby capturing new "knowledge" about the concepts  
15 embodied in the information system 125 and in the user's choice of queries (110).

Once the concept dictionary 140 has been established through a period of use of the apparatus 100, it may be used by the query editor and generator 105 to enable a user to select further queries to use in interrogating the information system 125 according to whether the user wishes to expand the scope of information retrieval, to  
20 reduce its scope or merely to search the information system 125 using semantically equivalent queries. Each time the user does use the apparatus 100 to retrieve information, particularly when the user enters a new query 110 not previously used, the knowledge acquisition module 135 is able to constantly update and improve the store of "knowledge" in the concept dictionary 140 for the ongoing benefit of users of the  
25 information system 125.

In a preferred embodiment of the present invention, to be described below, the knowledge acquisition module 135 and the concept dictionary 140 are arranged, respectively, to process and to store fuzzy relationships between queries and hence to provide a less precise (less "crisp") and thus more appropriate measure of semantic  
30 equivalence for storage in the concept dictionary 140. This has the advantage that lines of enquiry may be suggested to and selected by users of the apparatus 100 that would not ordinarily have been apparent with more precise "crisp" processing, with the potential to yield more useful results from the information system 125. The decision to use fuzzy processing techniques in preferred embodiments of the present invention  
35 recognises the fact that information retrieval on the basis of user-supplied queries is a

relatively imprecise process. Fuzzy processing has the potential to extract more useful information from the implicit and explicit assumptions behind a user's choice of input query and the body of information in the information system 125 than is possible with crisp processing of semantic relationships.

- 5 However, before discussing the preferred use of fuzzy processing by the apparatus 100, an example will be described to show how the concept dictionary 140 may be populated with "knowledge" acquired using "crisp" processing techniques.

Consider two queries  $Q_1$  and  $Q_2$ , with their corresponding answer sets  $S_1$  and  $S_2$  obtained by interrogating the information system 125. Assume these answer sets to  
10 be completely certain, rather than weighted to some degree of relevance. Assume that

$$Q_1 = \text{"find a garage in Ipswich"}$$

and

$$Q_2 = \text{"find car repair in Ipswich"}$$

and that the information system 125 returns a set of answers to the second query,  $S_2$ ,  
15 which is a subset of  $S_1$ . It may be deduced from this that "car repair" is a term having a more restricted meaning than the term "garage". A human expert is able to recognise cases of generalisation and specialisation in queries, but known techniques can also be used to achieve this automatically, for example with reference to a lexical database such as Wordnet™, accessible over the Internet at <http://www.cogsci.princeton.edu/~wn/>, and  
20 able to supply, for example, *hotel* as a synonym for the noun *inn*. If this is a valid equivalence in the context of the information system 125, it may be expected that the information system 125 would return an identical sets of answers in response to a query searching for hotels in a particular location and to a query searching for inns in the same location.

25 Formally, let  $Q(x)$  denote a query predicate that returns *true* or *false* according to whether or not an entry  $x$  is relevant to the query  $Q$ . Then the set of solutions

$$SQ = \{x \mid Q(x)\}$$

is the set of all entries  $x$  that satisfy (are relevant to) the query  $Q$ . It can be stated that for two queries,  $Q$  and  $P$ :

30  $Q$  generalises  $P$  if  $SP \subseteq SQ$

$Q$  specialises  $P$  if  $SQ \subseteq SP$

$Q$  is equivalent to  $P$  if  $SQ = SP$

Consider the following set of queries and corresponding answers:

<i>id</i>	<i>queryid</i>	<i>Query</i>	<i>ansid</i>	<i>Answer Entry</i>				
1	q1	car hire in Ipswich	a1	Eurodollar rent a car				
2	q1	car hire in Ipswich	a2	Autorent (UK)				
3	q2	car rental in Ipswich	a1	Eurodollar rent a car				
4	q2	car rental in Ipswich	a2	Autorent (UK)				
5	q3	restaurant in Suffolk	a3	Church Yards Seafood	Restaurant			
6	q3	restaurant in Suffolk	a4	Curry Inn				
7	q3	restaurant in Suffolk	a5	Passage To India				
8	q3	restaurant in Suffolk	a6	Chicago Rock Cafe				
9	q4	restaurant in Ipswich	a5	Passage To India				
10	q4	restaurant in Ipswich	a6	Chicago Rock Cafe				
11	q5	Indian restaurant in Suffolk	a4	Curry Inn				
12	q5	Indian restaurant in Suffolk	a5	Passage To India				
13	q6	Indian restaurant in Ipswich	a5	Passage To India				

By the above reasoning, the answers to queries q3, q4, q5 and q6 in the table above can be used to deduce that *Ipswich* is more specific term than *Suffolk* and that *Indian Restaurant* is a more specific term than *Restaurant*. Such deduced information may 5 be stored in a concept dictionary 140 and used subsequently to assist users in generalising or specialising their queries.

- The relationships between queries or query terms as derived in the example above are examples of "crisp" relationships. They are derived on the basis that the answers to the submitted queries are certain. In practice this is not generally the case.
- 10 The preferred approach for use in embodiments of the present invention is to extend the ideas above to allow partial relevance of answer entries to queries and to convert the crisp relationships into fuzzy relationships. In this preferred approach the definitions of generalisation, specialisation and equivalence are expanded to cater for partial inclusion and approximate equality.
- 15 A method will now be described for deriving relationships between queries using fuzzy processing techniques for implementation by the apparatus 100 and in particular by the knowledge acquisition module 135 according to a preferred embodiment of the present invention. Preferably, the knowledge acquisition module 135 determines the degrees to which a query *P* generalises a query *Q* and to which the query *P* specialises

the query Q for each pair of queries P and Q, in the context of the information system 125, using a representation framework known as the "mass assignment framework" in combination with a technique for calculating conditional probabilities of fuzzy sets called "semantic unification". These techniques are taught for example in the following published documents: J. F. Baldwin (1992) in "The Management of Fuzzy and Probabilistic Uncertainties for Knowledge-based Systems.", in the *Encyclopedia of AI*, edited by S. A. Shapiro, published by John Wiley (2<sup>nd</sup> edition), pages 528-537; J. F. Baldwin (1992) "Mass Assignments and Fuzzy Sets for Fuzzy Databases" in *Advances in the Shafer Dempster Theory of Evidence*, edited by M. Fedrizzi, J. Kacprzyk and R. R. Yager, published by John Wiley; J. F. Baldwin and T. P. Martin (2001) in "Towards Inductive Support Logic Programming", *Proc. IFSA-NAFIPS 2001*, Vancouver, pages 1875-1880; and J. F. Baldwin, J. Lawry, and T. P. Martin in "Efficient Algorithms for Semantic Unification", in *Proc. Information Processing and the Management of Uncertainty*, 1996, Spain.

Considering firstly a proposition that a query P generalises a query Q. This proposition is represented by the rule

$$\text{Relevant}(P, E) \leftarrow \text{Relevant}(Q, E)$$

where E is an entry (set of information) identifiable in the information system 125. The degree to which this rule applies in respect of the queries P and Q may be calculated from the fuzzy conditional

$$20 \quad \{x : \text{Relevant}(P, x)\} \mid \{x : \text{Relevant}(Q, x)\}$$

where x is a set of information in the information system 125, the calculation being performed over mass assignment elements making up fuzzy answer relations SP and SQ. For example, suppose that execution of the query P by the information retrieval module 120 returns the fuzzy answer relation

$$25 \quad \text{SP} = \{\text{a1} : 1, \text{a2} : 1, \text{a3} : 0.7, \text{a4} : 0.6\}$$

and execution of the query Q returns

$$\text{SQ} = \{\text{a1} : 1, \text{a2} : 0.8, \text{a3} : 0.5\}$$

In these fuzzy answer relations, a1, a2, ... etc. are answer identifiers, e.g. as used in the table above, and the values are fuzzy membership values for each answer calculated for example by the information retrieval module 120 by conventional means and representative of the degree to which the respective answer would be included in a response to the respective query by the information system 125. Each value is essentially a measure of the relevance of the answer to the query as may be determined by any one of a number of known information retrieval algorithms.

Intuitively, from an inspection of the fuzzy answer relations SP and SQ, the query P seems to be more general than the query Q, since Q returns fewer answers from the information system 125 and lower membership values in two cases (a2 and a3) than P.

To calculate the degree of support for the proposition that the query Q is generalised by P,

- 5 a mass assignment is firstly formed on each of the fuzzy answer relations, as follows:

$$m(SP) = \{\langle a1, a2 \rangle : 0.3, \langle a1, a2, a3 \rangle : 0.1, \langle a1, a2, a3 \rangle, \\ \langle a1, a2, a3, a4 \rangle : 0.6\}$$

10  $m(SQ) = \{\langle a1 \rangle : 0.2, \langle a1 \rangle, \langle a1, a2 \rangle : 0.3, \langle a1 \rangle, \langle a1, a2 \rangle, \langle a1, a2, a3 \rangle : 0.5\}$

where the notation

$$\{\langle a1 \rangle, \langle a1, a2 \rangle\} : 0.3$$

- indicates a degree of support of 0.3, from an interval [0,1], for the set of relevant answers  
 15 to be either a1 or both a1 and a2, the values, e.g. 0.3, being obtained by subtracting consecutive fuzzy membership values in the fuzzy relations SP and SQ. For example, in the mass assignment for SP, answer a1 cannot arise in isolation because the answer a2 also has a fuzzy membership value of 1 in the fuzzy relation SP, so the probability mass for  $\{\langle a1 \rangle\}$  is 0. However, the probability mass for  $\{\langle a1, a2 \rangle\}$  is  $1-0.7=0.3$ , and that for  
 20  $\{\langle a1, a2 \rangle, \langle a1, a2, a3 \rangle\}$  is  $0.7-0.6=0.1$ , etc.

The next step is to use the "standard point semantic unification" algorithm, described for example in the last of the four references listed above, to derive the degree of support for the rule

$$Relevant(P, E) \leftarrow Relevant(Q, E)$$

- 25 from the mass assignments  $m(SP)$  and  $m(SQ)$ .

- For each of the answer combinations arising for the query Q, the question to be asked in the semantic unification process is: is it possible, and if so what is the probability that given a particular answer combination for the query P, the answer combination for Q would arise? The answers to this question are presented for each of the queries in the  
 30 table below, where the mass assignments for SQ are written along the top of the table and those for SP are written down the left hand side.

	$\{\langle a1 \rangle : 0.2$	$\{\langle a1 \rangle, \langle a1, a2 \rangle : 0.3\}$	$\{\langle a1 \rangle, \langle a1, a2 \rangle, \langle a1, a2, a3 \rangle : 0.5\}$
$\{\langle a1, a2 \rangle : 0.3\}$	0	$1/2 \times 0.3 \times 0.3$	$1/3 \times 0.3 \times 0.5$
$\{\langle a1, a2 \rangle, \langle a1, a2, a3 \rangle : 0.1\}$	0	$1/2 \times 0.1 \times 0.3$	$2/3 \times 0.1 \times 0.5$
$\{\langle a1, a2 \rangle, \langle a1, a2, a3 \rangle, \langle a1, a2, a3, a4 \rangle : 0.6\}$	0	$1/2 \times 0.6 \times 0.3$	$2/3 \times 0.6 \times 0.5$

Taking the first column, first row, it can be seen intuitively there is no possibility that if the answer to a query was  $\langle a1 \rangle$  alone that the answer to the query can be  $a1$  and  $a2$  ( $\langle a1, a2 \rangle$ ). However, in the first row, second column the question asked is whether the 5 answer could be  $\{\langle a1, a2 \rangle, \langle a1, a2, a3 \rangle\}$  given that it was  $\{\langle a1 \rangle, \langle a1, a2 \rangle\}$ . The probability of this is the product of the individual probability masses multiplied by a factor indicative of the likelihood of the common answer combinations arising within the given answer combination. In the case of the first row, second column, assuming  $\langle a1 \rangle$  and  $\langle a1, a2 \rangle$  to be equally likely gives the factor 1/2 since if the answer is  $\langle a1, a2 \rangle$  then the answer could 10 be  $\{\langle a1, a2 \rangle, \langle a1, a2, a3 \rangle\}$  whereas if the answer is  $\langle a1 \rangle$  then it cannot. Each cell is weighted by the corresponding likelihood factor and the product of the respective probability masses, and the overall degree of support (semantic unification value) for the rule

*Relevant(P, E)  $\leftarrow$  Relevant(Q, E)*

15 is calculated as the sum over all cells in the table, giving a semantic unification value for this rule of 0.433.

A similar exercise can be carried out to test the support for the rule

*Relevant(Q, E)  $\leftarrow$  Relevant(P, E)*

which gives for this example a semantic unification value 0.548.

20 The knowledge acquisition module 135 is arranged to perform the fuzzy analysis described above in respect of each combination of queries selected from the input query 110 and the corresponding queries generated by the query editor and generator 105, using the corresponding answer responses obtained from the information system 125 by the query execution and information retrieval module 120. The semantic unification values 25 representing the degree of support for generalisation and for specialisation of one query

by another are calculated and stored by way of an update to the concept dictionary 140, along with the respective queries themselves (if not already stored).

The process of updating the concept dictionary 140 starting from receipt of an input query 110 can be summarised and will now be described with reference to the flow 5 diagram of Figure 2.

Referring to Figure 2, and additionally to Figure 1, at STEP 200 an input query 110 is received by the query editor and generator 105 in the apparatus 100. At STEP 205, the query editor and generator 105 generates a set of queries related semantically to the input query 110 with reference to an external lexical reference source 115 such as 10 Wordnet, referenced above. In particular, the external lexical reference source 115 is used to obtain, for a noun of the input query 110, at least one of three types of semantically related noun, as follows (these are WordNet options, for example):

- o Synsets - roughly equivalent terms
- 15 o Hypernym - super types (less restricted terms)
- o Hyponym - subtype (more restricted terms)

Each of the returned nouns is used to generate a related query by replacing the respective noun in a copy of the input query 110.

20 At STEP 210, the input query 110 and the related queries generated at STEP 205 are executed by the information retrieval module 120 to identify sets of information stored in the information system 125 relevant to each of those queries. Preferably, in order to distinguish one set of information identified by the information retrieval module 120 from another, where distinct identifiers are not already defined and returned for each 25 different set of information in the information system 125, then either the information retrieval module 120 itself or the knowledge acquisition module 135 are arranged to compare retrieved sets of information and to assign a unique identifier to each distinct set of information so identified. If a particular related query returns no answers, it is assumed to be an inappropriate change to the input query 110 and is discarded.

30 Those sets of information identified as being relevant to the input query 110 in particular, or at least assigned identifiers and/or references to those sets of information, are output at STEP 215 as a set of output answers 130 in response to the input query 110. At STEP 220 the information retrieval results output from the module 120, following execution of the queries at STEP 210, are analysed by the knowledge acquisition module 35 135, along with the input query 110 and the related queries from STEP 205, to determine

- the degree of support for each of the different semantic relationships, i.e. generalisation, specialisation and hence similarity, between those queries, using one of the methods described above. The results of this analysis are used at STEP 225 to derive new knowledge about the information system 125, in particular to deduce the position of a
- 5 newly input or derived query or its constituent terms in a semantic hierarchy of queries and/or query terms, and to analyse the queries themselves to deduce whether a particular term of one query is semantically equivalent or related by generalisation or specialisation to a term of another query. In order to deduce semantic equivalence, the knowledge acquisition module 135 is arranged to interpret the semantic unification values associated
- 10 with generalisation and specialisation of one query by another: for example, if both values are "high", the respective queries are interpreted to be semantically equivalent and a value representative of the degree of equivalence is taken to be the mean of the generalisation and specialisation values; if the value for generalisation is "low" and that for specialisation is "high", or vice versa, then specialisation or generalisation, respectively,
- 15 holds; and if both values are "low", the semantic relationship between the queries is considered "weak". A threshold value or a fuzzy set may be defined by the knowledge acquisition module 135 in respect of the information system 125, to control interpretation of "high" and "low". For example, "low" may be a value "below 0.5", or a fuzzy set may define "low" as {< 0.3 is definitely low, >0.5 is definitely not low, 0.3-0.5 is fuzzy low}.
- 20 However, the value ranges applicable in respect of a particular information system 125 may be adjusted by means of simple experiments.

At STEP 230, the results of the analysis step 220 and the derive new knowledge step 225 are used to update a concept dictionary 140 generated and maintained by the apparatus 100 in respect of the information system 125.

- 25 Preferably, the concept dictionary 140 comprises data representative of a graph structure having nodes comprising query words or terms, e.g. "garage", interlinked, where respective relationships have been derived, with the respective values indicating the degree of support calculated for the relationship – generalisation, specialisation or equivalence. The links represented in the concept dictionary 140 may be followed from
- 30 one node to another to obtain a more generalised or more specialised word or phrase. Each link is a two-way link; following the link in one direction leads to a semantically more specialised node, in the other direction to a more general node, in the context of the respective information system 125. Preferably, a hash table is stored as part of the concept dictionary 140 to provide a link to a node of the graph structure from a given word
- 35 or phrase, e.g. one entered by a user at a user interface. By way of example, a portion of

a graph structure represented by data stored in the concept dictionary 140 will now be described with reference to Figure 3.

Referring to Figure 3, a graph structure is shown comprising a number of query nodes 300-330 and links therebetween representative of derived semantic relationships.

- 5 In particular, the query node 300 "*garage in Ipswich*" is shown linked to the query node 305 "*buy car in Ipswich*". Stored semantic unification values, calculated in respect a specialisation of the query node 300 by the query node 305, and vice versa, are also shown alongside the links 335 and 340 respectively. Also shown as part of each of the query nodes 305-330 are statements 345-370 derived during STEP 225 of the process
- 10 described above with reference to Figure 2. These statements defining the strength of relationships between query terms, measures (in the range [0,1]) of the degree of "similarity" between terms in the example of Figure 3. For example, in query node 305, it has been calculated, using the semantic unification vales derived in respect of the relationship between query node 305 and query node 300, that the term "*buy car*" is
- 15 similar to the term "*garage*" with fuzzy membership value 0.273. That is, the terms have been found to be relatively dissimilar, as would be expected given the semantic unification value of 0.835 in support of specialisation by the query node 305 of the node 300 and only 0.112 in support of specialisation by the query node 300 of the node 305.

When presented in the form shown in Figure 3, the contents of the concept dictionary 140 can be seen to provide a useful source of information to users wishing to make alterations to queries for use in interrogating a respective information system 125. In particular, having received the results of a search of the information system 125 using a first query made up of terms already known in the concept dictionary 140, it would be clear from an inspection of links emanating from the node corresponding to the first query what alterations would need to be made to either generalise or specialise the first query to, respectively, expand or reduce the scope of the returned query results with a reasonable chance of success.

By way of example of the way in which a user may exploit the knowledge embodied in a concept dictionary 140 generated using preferred embodiments of the present invention, consider that the following knowledge has been accumulated in a concept dictionary 140, derived from previously used queries and query answers supplied by a respective information system 125, with each relationship having a high level of support (high semantic unification value):

*Italian restaurant generalisation\_of pizza*

35           *takeaway food generalisation\_of pizza*

takeaway food *generalisation\_of* fish and chips

takeaway food *generalisation\_of* Chinese takeaway

If a user finds that no answers are returned by the information system 125 in response to a query

5 "Find pizza in Ipswich"

then the knowledge (140) above may be used to suggest two possible query generalisations to improve the chances of obtaining useful answers, as follows:

"Find Italian restaurant in Ipswich"

"Find takeaway food in Ipswich"

10 If the user finds that the latter query was too general, i.e. it resulted in too many answers, then alternatives to this query may be offered, with reference to the knowledge above, by specialisation:

"Find fish and chips in Ipswich"

"Find Chinese takeaway in Ipswich"

15 In this way, not only has the user been able to make relevant adjustments to the choice of query in order to vary the responses given by the information system 125, but an alternative line of enquiry has also been suggested that may not have been apparent to the user of that particular information system 125.

Preferably, a user interface is provided with the apparatus 100 (not shown in  
20 Figure 1) to enable a user to submit queries 110 to the apparatus 100 and to receive output answers 130 from the apparatus 100 in response. The user interface may also be arranged to enable a user to navigate knowledge stored in the concept dictionary 140, preferably with the aid of a graphical user interface showing derived relationships between query nodes and query terms in a manner similar to that shown in Figure 3, in particular to  
25 enable the user to select particular queries and to request suggestions of more generalised, more specialised or semantically equivalent queries to execute in a respective information system 125.

Preferably, an apparatus 100 according to preferred embodiments of the present invention is implemented as a suite of computer programs using the Java programming  
30 language for running on a conventional server computer. The concept dictionary 140 is implemented using a conventional relational database management system such as Oracle™, although this too can be implemented using Java.

Besides use as an information retrieval method and apparatus, preferred embodiments of the present invention may be used to test the effectiveness of existing  
35 information retrieval systems. For example, the apparatus 100 may be linked to an

existing information retrieval system so that the query generator and editor 105 is arranged to receive (in a monitoring role) queries entered by a user of the existing system and the query execution and information retrieval module 120 is arranged with access to submit queries to the existing system and to receive corresponding answers. Over a 5 period of time in use, a concept dictionary 140 generated in respect of the existing system, by the process described above with reference to Figure 2, may be exported in a format useable in the existing system and used to test the effectiveness of a query interface provided by the existing system, for example by comparing the results of executing queries suggested by the existing system with the results of executing queries 10 suggested with reference to the generated concept dictionary 140.

In another mode of operation of preferred embodiments of the present invention, a bulk querying process may be implemented whereby a set of queries is built up and then sent into the apparatus 100 as input queries 110. This mode of operation may be particularly useful when a concept dictionary 140 needs to be generated quickly rather 15 than over an extended period of use of the apparatus 100 with a particular information system 125.

In another mode of operation of preferred embodiments of the present invention, a concept dictionary (140) generated in respect of a particular information system 125 may be exported in a format useable in another information retrieval system, also 20 arranged with access to the information system 125, as a source of knowledge for use in querying the information system 125 through the other information retrieval system.

## CLAIMS

1. A method of generating a concept dictionary (140) for use in querying an information system (125), comprising the steps of:
  - 5 (i) receiving an information search criterion;
  - (ii) deriving (105), using a lexical reference source (115), at least one search criterion having related meaning to said received search criterion (110);
  - (iii) identifying sets of information in said information system (125) relevant to said received search criterion (110) and to said at least one derived search criterion;
  - 10 (iv) analysing the identified sets of information to derive relationships between said received search criterion (110) and said at least one derived search criterion in the context of said information system (125); and
  - (v) storing, in a concept dictionary (140), information relating to said received (110) and said at least one derived search criterion and to respective said derived relationships
- 15 therebetween, for use in querying said information system (125).
  
- 20 2. A method as in Claim 1, wherein, at step (i), receiving an information search criterion (110) comprises selecting an information search criterion stored in said concept dictionary (140).
- 25 3. A method as in Claim 1 or Claim 2, wherein, at step (ii), deriving at least one search criterion having related meaning comprises replacing a term of said received search criterion (110) with a related term having a more specific meaning according to said lexical reference source (115).
- 30 4. A method as in any one of claims 1 to 3, wherein, at step (ii) deriving at least one search criterion having related meaning comprises replacing a term of said received search criterion (110) with a related term having a more general meaning according to said lexical reference source (115).
5. A method as in any one of claims 1 to 4, wherein, at step (ii) deriving at least one search criterion having related meaning comprises replacing a term of said received search criterion (110) with a related term having an equivalent meaning according to said lexical reference source (115).

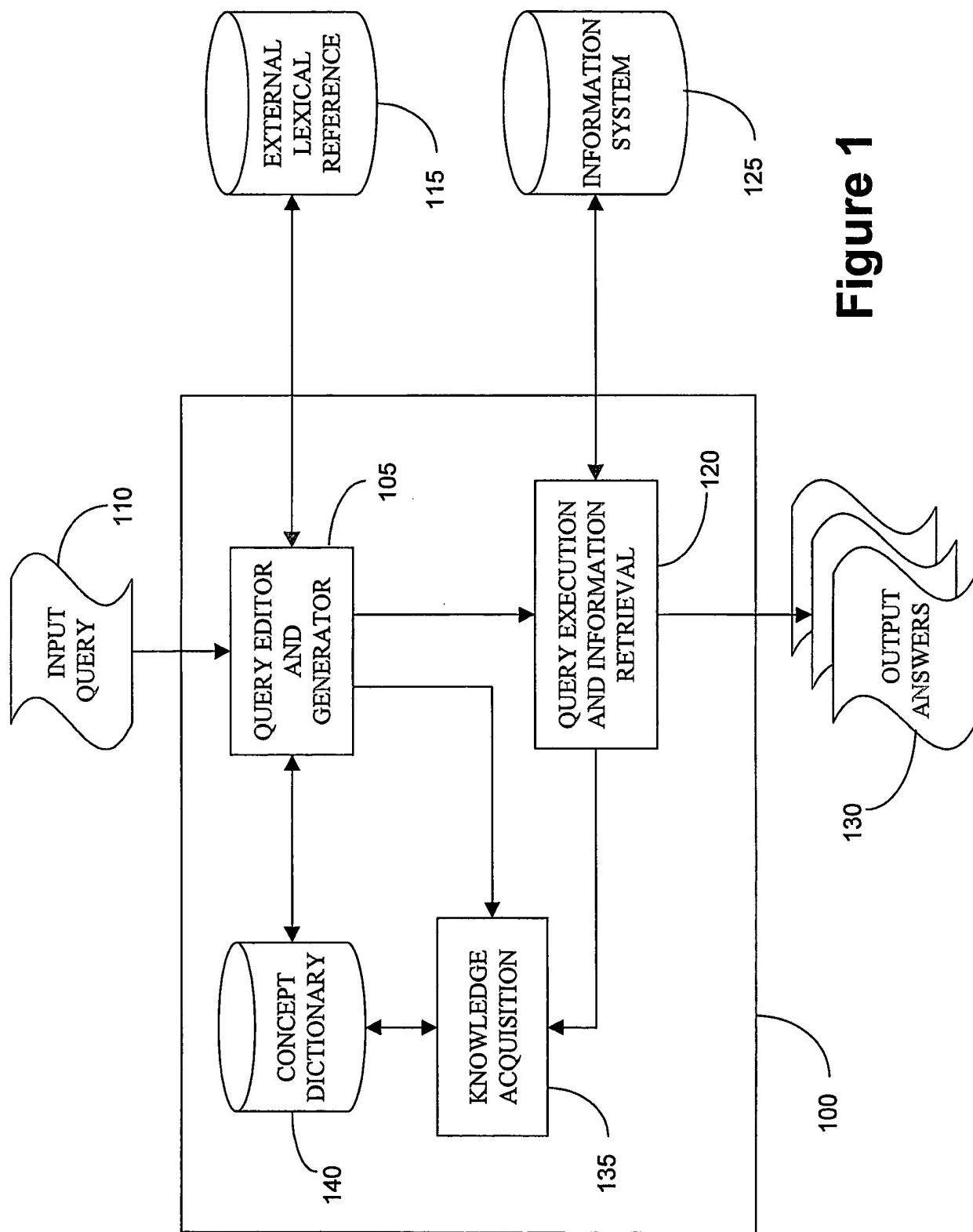
6. A method as in any one of the preceding claims, wherein, at step (ii), said lexical reference source (115) is a thesaurus.
7. A method as in any one of the preceding claims, wherein, at step (ii), said lexical reference source (115) is an ontological database.
8. A method as in any one of the preceding claims, wherein, at step (ii), a plurality of search criteria are derived, each having related meaning to said received search criterion (110), and wherein at step (iv), the respective identified sets of information are analysed to derive relationships between search criteria comprised in said plurality of derived search criteria.
9. A method as in any one of the preceding claims, wherein, at step (iv), deriving relationships between said search criteria comprises performing fuzzy processing of said derived search criteria and respective said identified sets of information to determine a measure of the generalisation and/or specialisation of one said search criterion over another in the context of said information system (125).
10. A method of accessing sets of information stored in an information system (125) using information search criteria stored in a concept dictionary (140) generated for the information system (125) according to the method in any one of claims 1 to 9, comprising the steps of:
  - (a) selecting a first information search criterion;
  - (b) using a search engine to identify one or more sets of information in the information system (125) relevant to said first information search criterion; and
  - (c) selecting at least one further information search criterion from search criteria stored in the concept dictionary (140), semantically related to said first information search criterion according to information stored in the concept dictionary (140), according to whether a more general, a more specialised or an equivalent search is required.
- 30 11. An information retrieval apparatus (100) for accessing sets of information stored in an information system (125), comprising;
  - an input for receiving an information search criterion (110);

deriving means (105) for deriving, using a lexical reference source (115), at least one search criterion having related meaning to said received information search criterion (110);

- retrieval means (120) for identifying sets of information in said information system
- 5 (125) relevant to said received search criterion (110) and to said at least one derived search criterion;
- analysis means (135) for analysing said identified sets of information to derive relationships between said received search criterion (110) and said at least one derived search criterion in the context of said information system (125); and
- 10 updating means for storing, in a concept dictionary (140), information relating to said received (110) and said at least one derived search criterion and to respective said derived relationships therebetween, for use in querying said information system (125).

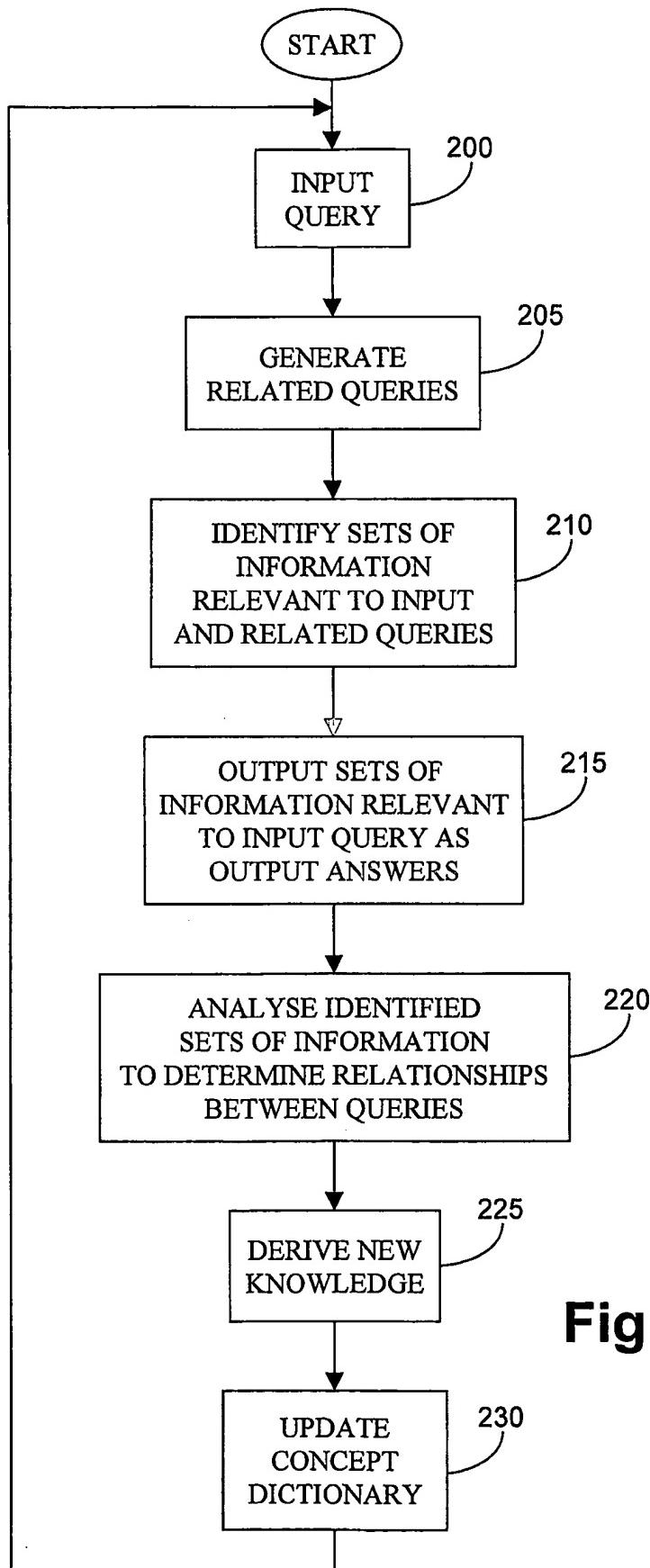
10/549365

1/3

**Figure 1**

**10/549365**

2/3

**Figure 2**

10/549365

3/3

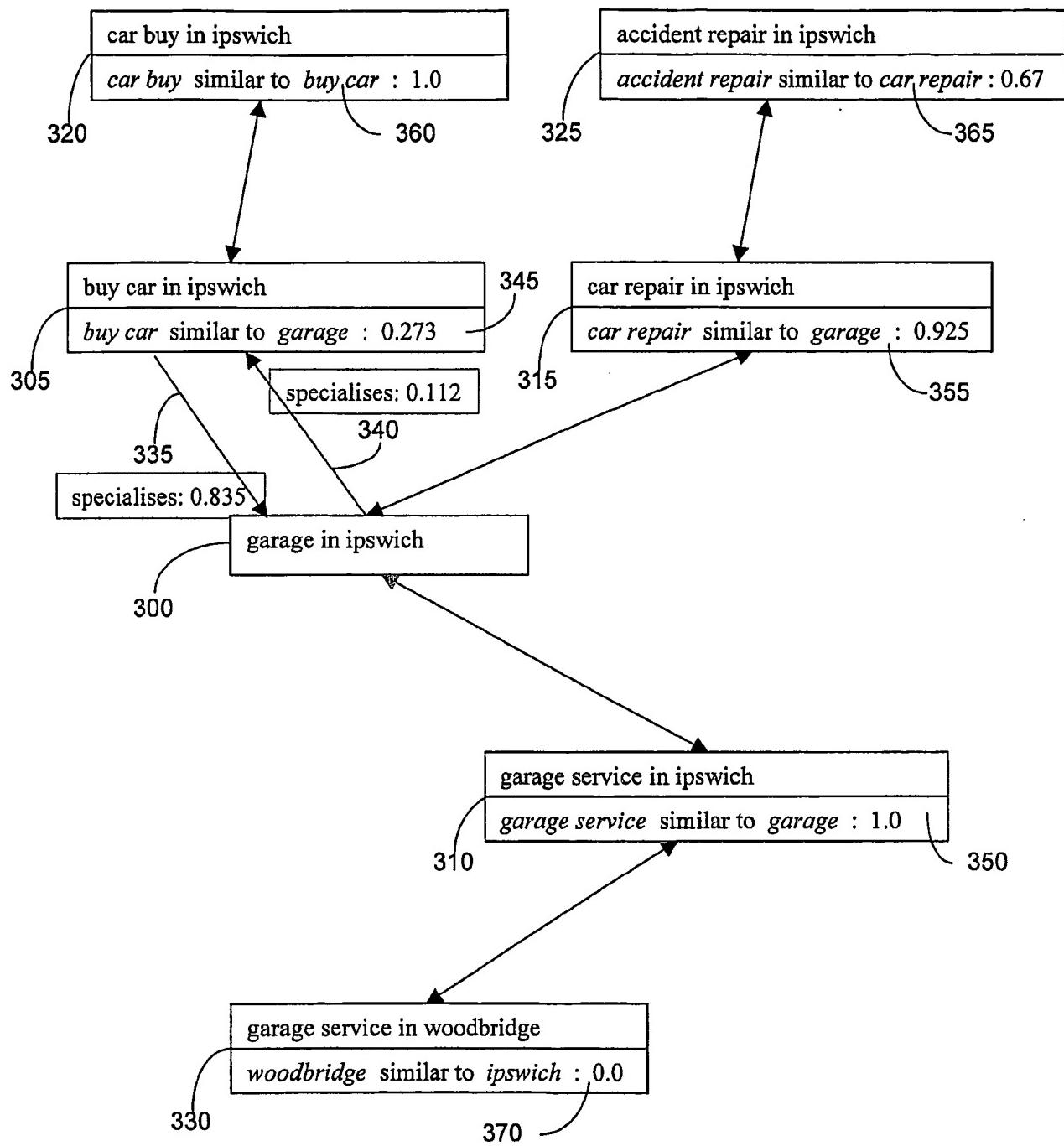


Figure 3

# INTERNATIONAL SEARCH REPORT

International Application No  
PCT/GB2004/000744

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)  
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the International search (name of data base and, where practical, search terms used)

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MENA E ET AL: "Observer: an approach for query processing in global information systems based on interoperation across pre-existing ontologies" COOPERATIVE INFORMATION SYSTEMS, 1996. PROCEEDINGS., FIRST IFCIS INTERNATIONAL CONFERENCE ON BRUSSELS, BELGIUM 19-21 JUNE 1996, LOS ALAMITOS, CA, USA, IEEE COMPUT. SOC P, US, 19 June 1996 (1996-06-19), pages 14-25, XP010200745 ISBN: 0-8186-7505-5 the whole document	1-8,11
Y	-----	9,10
X	WO 02/05137 A (CRITICALPOINT SOFTWARE CORP) 17 January 2002 (2002-01-17) page 1 - page 6; figures 6,18,27	1-8,11
Y	-----	9,10
		-/-

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

° Special categories of cited documents:

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the International filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the International filing date but later than the priority date claimed

- "T" later document published after the International filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed Invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed Invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- "&" document member of the same patent family

Date of the actual completion of the International search

16 June 2004

Date of mailing of the International search report

29/06/2004

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Eichenauer, L

**INTERNATIONAL SEARCH REPORT**

International Application No  
PCT/GB2004/000744

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MENA ET AL: "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies" DISTRIBUTED AND PARALLEL DATABASES, KLUWER, NL, vol. 8, no. 2, April 2000 (2000-04), pages 1-50, XP002201997 ISSN: 0926-8782 page 1 - page 36	1-8,11
Y		9,10
X	VISSEER ET AL: "Resolving Ontological Heterogeneity in the KRAFT Project" DATABASE AND EXPERT SYSTEMS APPLICATIONS, 1999. PROCEEDINGS. TENTH INTERNATIONAL WORKSHOP ON FLORENCE, ITALY 1-3 SEPT. 1999, LOS ALAMITOS, CA, USA, IEEE COMPUT. SOC, US, 1 September 1999 (1999-09-01), pages 668-677, XP002201998 ISBN: 0-7695-0281-4	1-8
Y	the whole document	9,10
Y	TAKAGI T ET AL: "Realization of sound-scape agent by the fusion of conceptual fuzzy sets and ontology" FUZZY SYSTEMS CONFERENCE PROCEEDINGS, 1999. FUZZ-IEEE '99. 1999 IEEE INTERNATIONAL SEOUL, SOUTH KOREA 22-25 AUG. 1999, PISCATAWAY, NJ, USA, IEEE, US, 22 August 1999 (1999-08-22), pages 801-806, XP010350907 ISBN: 0-7803-5406-0 the whole document	9,10
Y	CROSS V V ET AL: "Fuzzy ontologies for multilingual document exploitation" NORTH AMERICAN FUZZY INFORMATION, 1999. 18TH INTERNATIONAL CONFERENCE OF THE, NAFIPS NEW YORK, NY, USA 10-12 JUNE 1999, PISCATAWAY, NJ, USA, IEEE, US, 10 June 1999 (1999-06-10), pages 392-397, XP010343038 ISBN: 0-7803-5211-4 the whole document	9,10
Y	GRANTNER J L ET AL: "Hybrid fuzzy-Boolean automata for ontological controllers" FUZZY SYSTEMS PROCEEDINGS, 1998. IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE., THE 1998 IEEE INTERNATIONAL CONFERENCE ON ANCHORAGE, AK, USA 4-9 MAY 1998, NEW YORK, NY, USA, IEEE, US, 4 May 1998 (1998-05-04), pages 400-404, XP010287483 ISBN: 0-7803-4863-X the whole document	9,10

**INTERNATIONAL SEARCH REPORT**

Information on patent family members

national Application No  
PCT/GB2004/000744

Patent document cited in search report	Publication date		Patent family member(s)	Publication date
WO 0205137	A 17-01-2002	AU WO US	7189101 A 0205137 A2 2002059289 A1	21-01-2002 17-01-2002 16-05-2002